Requested Patent: JP2000048001A

Title: REAL-TIME SHARED DISK SYSTEM FOR COMPUTER CLUSTERS ;

Abstracted Patent: US6182197 ;

Publication Date: 2001-01-30 ;

Inventor(s): MUKHERJEE RAJAT (US); DIAS DANIEL MANUEL (US) ;

Applicant(s): IBM (US) ;

Application Number: US19980113752 19980710 ;

Priority Number(s): US19980113752 19980710 ;

IPC Classification: G06F12/00 ;

Equivalents: SG77258

ABSTRACT:

A clustered computer system includes a shared data storage system, preferably a virtual shared disk (VSD) memory system, to which the computers in the cluster write data and from which the computers read data, using data access requests. The data access requests can be associated with deadlines, and individual storage devices in the shared storage system satisfy competing requests based on the deadlines of the requests. The deadlines can be updated and requests can be killed, to facilitate real time data access for, e.g., multimedia applications such as video on demand .

(19)日本国特許庁（ＪＰ）　(12)公開特許公報(A)　(11)特許出願公開番号
特開2000－48001
(P2000－48001A)
(43)公開日　平成12年２月18日(2000.2.18)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
G06F	15/16	640	G06F	15/16	640A
	12/00	535		12/00	535M
	15/177	682		15/177	682G

審査請求　未請求　請求項の数24　ＯＬ　（全 11 頁）

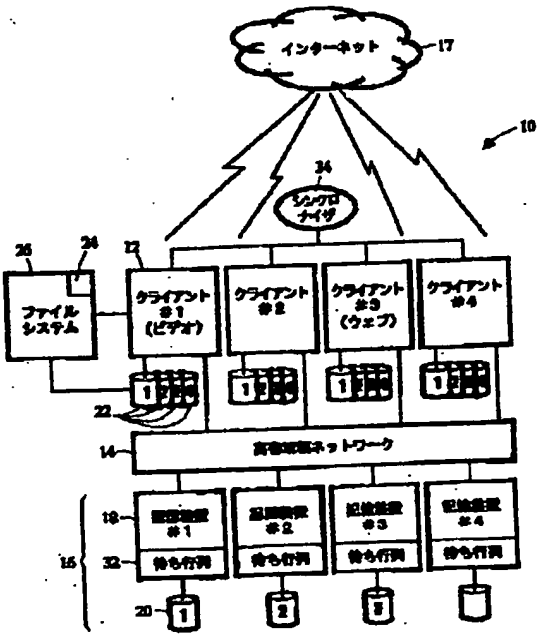
(21)出願番号　特願平11－182880
(22)出願日　平成11年６月29日(1999.6.29)
(31)優先権主張番号　０９／１１３７５２
(32)優先日　平成10年７月10日(1998.7.10)
(33)優先権主張国　米国（US）

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
INTERNATIONAL BUSINESS MASCHINES CORPORATION
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）
(72)発明者　ダニエル・マヌエル・ディアス
アメリカ合衆国10541　ニューヨーク州マホパック　バイク・プレース　16
(74)代理人　100086243
弁理士　坂口　博　（外１名）

最終頁に続く

(54)【発明の名称】　コンピュータ・クラスタ用のリアルタイム共用ディスク・システム

(57)【要約】
【課題】　クラスタ化されたコンピュータ・システム内の共用記憶システムに対して、優先順位付きのデータアクセス要求の発行、更新および打切りを行うための、汎用コンピュータを提供すること。
【解決手段】　クラスタ化されたコンピュータ・システムは、共用データ記憶システム、好ましくは仮想共用ディスク（ＶＳＤ）記憶システムを含み、この共用データ記憶システムに対して、クラスタ内のコンピュータが、データ・アクセス要求を使用してデータを書き込み、データを読み取る。データ・アクセス要求に、締切期限を関連付けることができ、共用記憶システム内の個々の記憶装置が、要求の締切期限に基づいて競合する要求を満足する。たとえばビデオ・オン・デマンドなどのマルチメディア・アプリケーションのための、リアル・タイム・データ・アクセスを簡単にするために、締切期限を更新することができ、要求を切ることができる。



【特許請求の範囲】
【請求項１】１つまたは複数のデータ・アクセス要求にそれぞれの優先順位を関連付けるための論理手段と、
上記データ・アクセス要求および優先順位を記憶ノードに送るための論理手段と、
上記データ・アクセス要求が、それぞれの優先順位を考慮して満足されるように、上記それぞれの優先順位に基づいて上記記憶ノードで上記データ・アクセス要求を順序付けるための論理手段とを含む、１つまたは複数の記憶ノードにデータ・アクセス要求を通信する複数のクライアント・ノードを含むコンピュータ・システム。
【請求項２】更新された優先順位にするために、少なくとも１つのデータ・アクセス要求の優先順位を、記憶ノードによって上記データ・アクセス要求が満たされる前に変更するための論理手段と、
上記更新された優先順位に基づいて、上記記憶ノードでデータ・アクセス要求を再順序付けするための論理手段とをさらに含む、請求項１に記載のシステム。
【請求項３】さらに、少なくとも１つのデータ・アクセス要求を打ち切るための論理手段を含む、請求項１に記載のシステム。
【請求項４】さらに、計算ノードおよび上記記憶ノードを互いに疎に同期化するための手段を含む、請求項１に記載のシステム。
【請求項５】各記憶ノードが、少なくとも１つの記憶コンピュータおよび少なくとも１つのデータ記憶装置を含み、上記記憶コンピュータが、上記データ記憶装置が上記データ記憶装置の内部の検討事項に基づいてデータ・アクセス要求に応答するシーケンスを再順序付けすることができなくなるように、上記データ記憶装置に対して１時に１つを超えないデータ・アクセス要求を送るための論理手段を含む、請求項１に記載のシステム。
【請求項６】上記システムが、仮想共用ディスク・システムである、請求項１に記載のシステム。
【請求項７】上記優先順位が、時間に基づく締切期限を含む、請求項１に記載のシステム。
【請求項８】共用記憶システムにデータ・アクセス要求を通信する複数のプロセッサを有するコンピュータ・システムにおいて、
データ記憶装置の外部の１つまたは複数の検討事項によって少なくとも部分的に定義される順序で上記データ・アクセス要求に応答するステップを含む、上記共用記憶システムの単一のデータ記憶装置に対する少なくとも２つの同時データ・アクセス要求を満足するための、コンピュータで実施される方法。
【請求項９】上記データ記憶装置の外部の上記１つまたは複数の検討事項が、データ要求優先順位を含む、請求項８に記載の方法。
【請求項１０】上記優先順位が、時間に基づく締切期限を含む、請求項９に記載の方法。
【請求項１１】１つまたは複数の上記データ・アクセス要求にそれぞれの優先順位を関連付けるステップと、
各記憶ノードが少なくとも１つのデータ記憶装置を含む、上記共用記憶システム内の記憶ノードに上記データ・アクセス要求および優先順位を送るステップと、
上記データ・アクセス要求が、それぞれの優先順位に従って満足されるように、上記それぞれの優先順位に基づいて上記記憶ノードで上記データ・アクセス要求を順序付けるステップとをさらに含む、請求項１０に記載の方法。
【請求項１２】更新された優先順位にするために、少なくとも１つのデータ・アクセス要求の優先順位を、上記データ・アクセス要求が記憶ノードによって満足される前に変更するステップと、
上記更新された優先順位に基づいて、上記記憶ノードでデータ・アクセス要求を再順序付けするステップとをさらに含む、請求項１１に記載の方法。
【請求項１３】少なくとも１つのデータ・アクセス要求を、上記要求が記憶ノードによって満足される前に打ち切るステップをさらに含む、請求項１０に記載の方法。
【請求項１４】計算ノードおよび上記記憶ノードを互いに疎に同期化するステップをさらに含む、請求項１０に記載の方法。
【請求項１５】各記憶ノードが、少なくとも１つの記憶コンピュータおよび少なくとも１つのデータ記憶装置を含み、上記記憶コンピュータが、上記データ記憶装置が上記データ記憶装置の内部の検討事項に基づいてデータ・アクセス要求に対する応答のシーケンスを再順序付けすることができなくなるように、上記データ記憶装置に対して１時に１つを超えないデータ・アクセス要求を送る、請求項１０に記載の方法。
【請求項１６】上記システムが、仮想共用ディスク・システムである、請求項８に記載の方法。
【請求項１７】ディジタル処理装置によって読み取ることができるコンピュータ・プログラム記憶装置と、
１つまたは複数のデータ・アクセス要求を満足するための方法ステップを実行するために上記ディジタル処理装置によって実行することのできる命令を含む、上記コンピュータ・プログラム記憶装置上のプログラム手段とを含み、上記方法ステップが、
上記データ・アクセス要求のうちの少なくとも一部に、それぞれの優先順位を関連付けるステップと、
共用記憶システムが上記優先順位を考慮して上記データ・アクセス要求に応答することができるように、上記優先順位および上記データ・アクセス要求を上記共用記憶システムに送るステップとを含む、コンピュータ・プログラム装置。
【請求項１８】上記共用記憶システムが、仮想共用ディスク・システムであり、上記優先順位のうちの少なくとも一部が、時間に基づく締切期限である、請求項１７に

記載のコンピュータ・プログラム装置。
【請求項１９】上記方法ステップが、さらに、
少なくとも１つのデータ・アクセス要求の優先順位を、更新された優先順位に基づいて上記共用記憶システム内で上記データ・アクセス要求を再順序付けできるようにするために、上記更新された優先順位にするために上記共用記憶システムによって上記データ・アクセス要求が満足される前に変更するステップを含む、請求項１７に記載のコンピュータ・プログラム装置。
【請求項２０】上記方法ステップが、さらに、
少なくとも１つのデータ・アクセス要求を、上記データ・アクセス要求が上記共用記憶システムによって満足される前に打ち切るステップを含む、請求項１９に記載のコンピュータ・プログラム装置。
【請求項２１】上記方法ステップが、さらに、上記データ・アクセス要求を互いに疎に同期化するステップを含む、請求項１７に記載のコンピュータ・プログラム装置。
【請求項２２】ディジタル処理装置によって読み取ることのできるコンピュータ・プログラム記憶装置と、
１つまたは複数のデータ・アクセス要求を満足するための方法ステップを実行するために上記ディジタル処理装置によって実行することのできる命令を含む、上記コンピュータ・プログラム記憶装置上のプログラム手段とを含み、上記方法ステップが、
それぞれの優先順位に基づいて、上記データ・アクセス要求のうちの少なくとも一部に対して、メモリ・システムを用いて応答するステップを含み、上記優先順位および上記データ・アクセス要求が、上記メモリ・システムに送られる、コンピュータ・プログラム装置。
【請求項２３】上記メモリ・システムが、共用記憶システムであり、上記優先順位が、時間に基づく締切期限であり、上記共用記憶システムが、複数の記憶ノードを含み、各記憶ノードが、少なくとも１つの記憶コンピュータおよび少なくとも１つのデータ記憶装置を含み、上記方法ステップが、上記データ記憶装置の内部の検討事項に基づいてデータ・アクセス要求に対する応答のシーケンスを上記データ記憶装置が再順序付けすることができなくなるように、上記データ記憶装置に対して１時に１つを超えないデータ・アクセス要求を送るステップを含む、請求項２２に記載のコンピュータ・プログラム装置。
【請求項２４】上記方法ステップが、変更された優先順位メッセージに応答して、上記共用記憶システムによってデータ・アクセス要求が満足される前に、記憶コンピュータ側で上記データ・アクセス要求を再順序付けするステップを含む、請求項２３に記載のコンピュータ・プログラム装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、全般的にはクラスタ化されたコンピュータ・システムに関し、具体的には、クラスタ化されたコンピュータ環境で共用ディスク・データにアクセスするためのシステムおよび方法に関する。
【０００２】
【従来の技術】クラスタ化されたコンピューティングとは、多数のコンピュータ・プロセッサが互いに調和して働いて、大規模な計算タスクのそれぞれの部分を引き受けるか、共通のデータ記憶資源を使用して別々のアプリケーションを実行する能力を指す。コンピュータは、ネットワークを介して互いにまたはデータ記憶資源と通信して、たとえば、計算の難しい仕事をコンピュータ間で分散したり、それぞれ独立のアプリケーションを実行する。一緒に働いて大規模な計算を引き受ける多数のコンピュータを使用することは、しばしば、そのような作業を実行するために単一のモノリシック・プロセッサを提供することよりコスト効率が良い。
【０００３】クラスタ化されたコンピューティング・システムでは、各コンピュータは、通常はコンピュータ・ディスクである１つまたは複数のそれぞれのデータ記憶装置に物理的に接続される可能性がある。さらに、複数のコンピュータがディスク上のデータにアクセスできるようにするために、ディスクは、複数のコンピュータに物理的に接続される可能性がある。このような共用システム・ディスクは、システムのコンピュータに物理的に接続されるので、「物理的に共有される」ディスクと称する。共用記憶システムと称するこのようなシステムでは、複数のコンピュータの間で計算が分散されるだけではなく、複数のディスクにまたがってデータも分散されることが容易に理解される。
【０００４】複数のコンピュータへのディスクの物理的接続は、記憶装置を共有するための効率的な方法であるが、それに付随する入出力ハードウェアが、相対的に高価になる場合がある。さらに、このようなシステムは、簡単にはスケーリングできない。具体的に言うと、多数のコンピュータのそれぞれを多数のディスクのそれぞれに接続することは、高価であるだけではなく、過多な複雑な配線を必要とする。
【０００５】したがって、本発明の譲受人は、「仮想共用ディスク」または「ＶＳＤ」と称するシステムを導入した。ＶＳＤにおいては、クラスタ化されたコンピューティング・システムの各コンピュータが、多数のシステム・ディスクのそれぞれを、各ディスクが物理的には１つまたは少数のコンピュータだけに接続されているという事実にかかわらず、すべてのシステム・コンピュータに物理的に接続されているとみなす。ＶＳＤでは、基本的に、コンピュータごとに、それぞれのシステム・ディスクを表すソフトウェア・モジュールと、コンピュータにとってそれぞれのディスクのデバイス・ドライバに見

えるソフトウェア・モジュールを提供することによって、これが達成される。物理的に接続されていないディスクを読み書きするためには、コンピュータは、物理的に共用されるディスクの場合のように「仮想」ディスクのデバイス・ドライバを呼び出して、読取動作または書込動作を要求し、その下層のVSDソフトウェア・モジュールが、そのディスクに実際に物理的に接続されているシステム内のコンピュータにその要求を送る。

【0006】しかし、共用システム・ディスクが仮想ディスクであるか物理ディスクであるかに無関係に、複数のシステム・コンピュータが、互いに事実上同時に単一のシステム・ディスクに読取要求または書込要求を発行する可能性があることが諒解される。これが発生した時には、システム・ディスクは、要求を受け取った順序や、ディスクの外部の検討事項に基づくのではなく、ディスクの内部の検討事項に基づいて、競合する要求に対処する。たとえば、システム・ディスクは、要求されたデータが記憶されている（またはこれから記憶される）ディスクのセクタに対するディスク・ヘッドの現在位置に基づいて、競合する要求に応答し、ヘッドに「最も近い」要求に最初に対処することができる。本明細書で使用される場合、そのような順序付けは、「内部」制約すなわち、ディスクの状態に依存し、要求の優先順位に依存しない制約に基づく。

【0007】したがって、本発明による認識に従えば、仮想的であれ物理的であれ1台のディスクに対する競合する要求に対処するための上述の処理は、1つの要求が別の要求より緊急性が高い可能性があるという事実を考慮していないという短所を有する。これは、通常は特定の時間的順序で配送されなければならない大量のデータ・ビットのブロックを読み取るマルチメディア・アプリケーションにおいて特に不利である。したがって、本明細書で認識されているように、読取要求と書込要求で、これを過ぎると要求が古くなり、その後に要求を満足しても結果的に無意味になる、応答締切期限（すなわち、時間に基づく優先順位）または他の時間に基づかない優先順位を定義することができる。しかし、上で述べたように、現在の共用記憶システムは、要求によって定義される優先順位に基づくのではなく、内部的なディスクの検討事項に基づいて要求に応答し、さらに、現在の共用記憶システムは、締切期限（または他の優先順位）以内に要求を満足できない場合にその要求を打ち切るかどうかを検討しない。

【0008】さらに、本発明では、あるマルチメディア・アプリケーションが、別のアプリケーションによって要求される可能性がある第2ビデオ・フレーム・ビットよりも先に再生されなければならない第1ビデオ・フレーム・ビットを要求する可能性があることが理解されている。このような状況では、共用記憶システムが、第1要求が第2要求の前に受け取られたかどうかに無関係に、第2要求の前に第1要求に応答することが望ましい。残念ながら、現在の共用ディスク・システムは、単一のシステム・ディスクに対する複数のほぼ同時の要求に応答する方法を決定する際に、競合する要求の相対的な優先順位の検討を考慮していない。

【0009】したがって、本発明では、VSDを含む共用記憶システムが、その共用記憶システム動作がリアルタイム・マルチメディア・データ・ストリーミングをモデル化するようにすることができる場合に、たとえばビデオオンデマンドなどのためのマルチメディア・ストリームのためのデータをよりよく供給することができることが認識されている。

【0010】

【発明が解決しようとする課題】本発明は、クラスタ化されたコンピュータ・システム内の共用記憶システムに対して、優先順位付きのデータアクセス要求の発行、更新および打切りを行うための、本明細書に記載の発明的ステップに従ってプログラムされた汎用コンピュータである。本発明は、ディジタル処理装置によって使用される、本発明を実施するためにディジタル処理装置によって実行可能な命令のプログラムを具体的に実施する、製造品（機械構成要素）として実施することもできる。本発明は、本明細書に記載の発明的方法ステップをディジタル処理装置に実行させる、クリティカルな機械構成要素で実現される。

【0011】

【課題を解決するための手段】本発明によれば、コンピュータ・システムに、複数の記憶ノードにデータ・アクセス要求を通信する複数のクライアント・ノードが含まれる。このシステムには、さらに、1つまたは複数のデータ・アクセス要求に、それぞれの優先順位、たとえば、時間に基づく締切期限を関連付けるための論理手段が含まれる。さらに、このシステムには、データ・アクセス要求と優先順位を記憶ノードに送るための論理手段が含まれる。さらに、このシステムには、データ・アクセス要求が、そのそれぞれの優先順位を考慮して満足されるように、それぞれの優先順位に基づいて記憶ノードでデータ・アクセス要求を順序付けするための論理手段が含まれる。

【0012】好ましい実施形態では、記憶ノードによって要求が満足される前に、データ・アクセス要求の優先順位を変更して更新された優先順位にするための論理手段が設けられる。論理手段は、その後、更新された優先順位に基づいて、記憶ノードでデータ・アクセス要求の再順序付けを行う。さらに、1つまたは複数のデータ・アクセス要求を打ち切るための論理手段を設けることができる。

【0013】計算ノードと記憶ノードは、互いに疎に同期化されることが好ましい。1実施形態では、システムが仮想共用ディスク（VSD）システムであり、各記憶

ノードに、少なくとも1つの記憶コンピュータと少なくとも1つのデータ記憶装置が含まれる。本発明がシステム・ディスク・コントローラで実施される場合を除いて、各記憶コンピュータには、データ記憶装置がデータ記憶装置の内部の検討事項に基づいてデータ・アクセス要求に対する応答のシーケンスを再順序付けすることができなくなるように、データ記憶装置に対して1時に1つを超えないデータ・アクセス要求を送るための論理手段が含まれる。しかし、ディスク・コントローラで実施される場合には、そのコントローラが、たとえばディスクの状態などの内部制約に基づく従来の要求順序付けと、本発明の外部順序付けを組み合わせることができる。

【0014】もう1つの態様では、共用記憶システムにデータ・アクセス要求を通信する複数のコンピュータを含むコンピュータ・システムにおいて、共用記憶システムの単一のデータ記憶装置に対する少なくとも2つの同時データ・アクセス要求を満足するための、コンピュータ実施される方法が開示される。この方法には、データ記憶装置の外部の1つまたは複数の検討事項によって定義される順序で要求に応答するステップが含まれる。

【0015】もう1つの態様では、コンピュータ・プログラム装置に、ディジタル処理装置によって読み取ることができるコンピュータ・プログラム記憶装置と、そのプログラム記憶装置上のプログラム手段とが含まれる。プログラム手段には、1つまたは複数のデータ・アクセス要求を満足するための方法ステップを実行するためにディジタル処理装置によって実行可能な命令が含まれる。下で詳細に開示するように、プログラム手段によって実施される方法ステップには、データ・アクセス要求のうちの少なくとも一部をそれぞれの優先順位と関連付けるステップと、その後、優先順位とデータ・アクセス要求を共用記憶システムに送るステップが含まれる。本発明を用いると、共用記憶システムは、優先順位を使用して要求に応答することができる。

【0016】もう1つの態様では、コンピュータ・プログラム装置に、ディジタル処理装置によって読み取ることができるコンピュータ・プログラム記憶装置と、プログラム記憶装置上のプログラム手段が含まれる。プログラム手段には、1つまたは複数のデータ・アクセス要求を満足するための方法ステップを実行するためにディジタル処理装置によって実行可能な命令が含まれる。下で詳細に開示するように、プログラム手段によって実施される方法ステップには、記憶システムを用いて、それぞれの優先順位に基づく順序で、データ・アクセス要求のうちの少なくとも一部に応答するステップが含まれ、優先順位およびデータ・アクセス要求が記憶システムに送られる。

【0017】

【発明の実施の形態】まず図1を参照すると、好ましくは高帯域幅ネットワークであるネットワーク14を介して、複数の記憶ノード16（開示を明瞭にするために4つの記憶ノード16だけが図示されている）を含む共用記憶システムにアクセスする、複数のクライアント・コンピュータ12（開示を明瞭にするためにクライアント・コンピュータ1ないし4だけが図示されている）を含む、全体的に符号10で示されるシステムが図示されている。「共用記憶」とは、共用される電子データ記憶装置を意味する。クライアント・コンピュータ12の可能な機能の1例として、図1で番号「1」を付されたクライアント・コンピュータ12は、ビデオ・サーバとすることができ、図1で番号「3」を付されたクライアント・コンピュータは、高速ワールド・ワイド・ウェブ・サーバ・コンピュータとすることができ、この両方のクライアント・コンピュータ12が、インターネットまたは他のWANなどの広域ネットワーク（WAN）17を介してそれぞれのクライアントと同時に通信する。

【0018】記憶ノード16のそれぞれには、少なくとも1つのそれぞれの記憶装置コントローラまたは記憶コンピュータ18と、たとえば関連する記憶コンピュータ18に物理的に接続されたデータ記憶ディスクとすることのできる少なくとも1つのデータ記憶装置20が含まれる。言い換えると、本発明は、符号18の要素が、下で呼称するように記憶コンピュータであるか、記憶装置コントローラであるかに無関係に、符号18の要素で実施することができる。記憶コンピュータ18は、クライアント・コンピュータ12と別に図示されているが、所望される場合には、クライアント・コンピュータ12のそれぞれが、そのクライアント・コンピュータに物理的に接続されたデータ記憶装置に関して、残りのクライアント・コンピュータに対してサーバ・コンピュータとして機能することができる。どの場合でも、本発明のコンピュータは、米国ニューヨーク州Armonkのインターナショナル・ビジネス・マシーンズ・コーポレーション（IBM）が製造するパーソナル・コンピュータなどのデスクトップ・コンピュータまたはラップトップ・コンピュータとすることができる。その代わりに、本発明のコンピュータは、IBM Network Stationsを伴うAS400などの商標の下で販売されるコンピュータ、UNIXコンピュータ、OS／2サーバ、Windows NTサーバ、IBM RS/6000 250ワークステーション、または他の同等の装置を含むコンピュータとすることができる。

【0019】好ましい実施形態では、システム10に、本発明の譲受人が所有し、参照によって本明細書に組み込まれる米国特許第5668943号明細書に開示された仮想共用ディスク（VSD）システムが使用される。したがって、前述の特許明細書に記載されているように、各クライアント・コンピュータ12は、仮想ディスク22がそのクライアント・コンピュータ12に物理的

に接続されているかのように、4台の仮想ディスク22と通信する。しかし、実際には、クライアント・コンピュータ12は、記憶装置に物理的に接続される必要はなく、図1で番号「1」を付された記憶コンピュータを、番号「1」を付されたディスクだけに物理的に接続することができる。同様に、番号「2」を付された記憶コンピュータを、番号「2」を付されたディスクだけに物理的に接続することができ、以下同様であり、クライアント・コンピュータ12とその仮想ディスクの間のアクセスが、上記特許明細書に従って管理される。しかし、本明細書に記載の原理は、一部またはすべてのシステム・ディスクが物理的に共用されるディスクである、すなわち、システム内の各クライアント・コンピュータに物理的に接続される、共用記憶システムにあてはまることを理解されたい。

【0020】本発明によれば、各クライアント・コンピュータ12は、締切期限モジュール24（開示を明瞭にするために図1では1つの締切期限モジュール24だけを図示する）にアクセスすることができる。締切期限モジュール24は、それぞれのクライアント・コンピュータ12によって実行するか、クライアント・コンピュータ12に関連するファイル・システム26（図1には1つのファイル・システム26だけを図示する）などのソフトウェア実施されるモジュールによって実行して、以下で詳細に説明する発明的論理を行うことができる。

【0021】締切期限モジュール24などの制御構成要素は、論理回路で実施されるものや、たとえばハード・ディスク装置または光ディスク装置などの、通常それぞれのクライアント・コンピュータ12に結合される適当な電子データ記憶装置に格納されたソフトウェアで実施されるものなどの論理構成要素によって実行されることを理解されたい。または、制御構成要素は、図2に示されたコンピュータ・ディスケット28などの他の論理構成要素で実施することができる。図2に示されたコンピュータ・ディスケット28は、コンピュータ可読コード手段（すなわちプログラム・コード要素）AないしDが格納されたコンピュータ使用可能媒体30を有する。どの場合でも、特定の論理構成要素またはコンピュータ・ソフトウェア・ステップの実行は、1つまたは複数のクライアント・コンピュータ12および記憶コンピュータ18の間で分散することができる。

【0022】本明細書の流れ図に、コンピュータ・プログラム・ソフトウェアとして実施された本発明の締切期限モジュールの構造を示す。当業者は、この流れ図が、コンピュータ・プログラム・コード要素または電子論理回路など、本発明に従って機能する論理要素の構造を示すことを諒解するであろう。明らかに、本発明は、その基本的な実施態様において、ディジタル処理装置（すなわちコンピュータ）に図示のステップに対応する一連の機能ステップを実行するように指示する形式で論理要素を表現する機械構成要素によって実行される。

【0023】言い換えると、締切期限モジュール24は、一連のコンピュータ実行可能命令として、関連するクライアント・コンピュータ12（またはファイル・システム26）内のプロセッサによって実行されるコンピュータ・プログラムとすることができる。上で述べた駆動装置に加えて、これらの命令は、たとえばコンピュータのRAMに常駐することができ、また、これらの命令は、DASDアレイ、磁気テープ、電子読取専用メモリ、または他の適当なデータ記憶装置に格納することができる。本発明の1実施形態では、コンピュータ実行可能命令を、コンパイルされたC++互換コードの行とすることができる。

【0024】以下の説明の後により明瞭になる理由のために、各記憶コンピュータ18には、データ読取要求およびデータ書込要求が優先順位によって順序付けされる、それぞれの要求待ち行列32が含まれる。符号18の要素がシステム・ディスク・コントローラでない場合には、関連するデータ記憶装置20には、1時に1つの入出力動作だけが提示される。また、ソフトウェア実施またはハードウェア実施される通常のタイム・シンクロナイザ34が、クライアント・コンピュータ12および記憶コンピュータ18のクロックを同期化することが好ましい。

【0025】「優先順位」とは、一般に、システム10がそれぞれの優先順位を使用して要求に応答するように制限する、データ・アクセス要求の制約を意味するが、システム10の状態に応じて、必ずしも要求の優先順位の絶対的な順序で要求が満足されるとは限らない。たとえば、上で述べたように、データ・アクセス要求は、ディスクの要求待ち行列内で優先順位によって順序付けされるが、特定の低優先順位の要求待ち行列が十分に短く、優先順位の低い要求、たとえば非リアル・タイム要求を先に満足し、その後に、優先順位の高い要求、たとえばリアル・タイム要求をその優先順位の制約（たとえば締切期限）内で満足することができる場合には、優先順位の低い要求、たとえば非リアルタイム要求が、後から到着したが優先順位の高い要求、たとえばリアル・タイム要求の前にサービスされる場合がある。さらに、ディスク・コントローラ上で実施される場合、コントローラは、関連するディスクの内部状態に基づく順序付けと、外部制約（すなわち要求の優先順位）を組み合わせることができる。

【0026】本明細書で使用する用語「優先順位」は、「サービスのクラス」を意味すると考えることもできることを諒解されたい。サービスのクラスという優先順位方式は、異なるサービスのクラスの要求が、必ずしも特定の順序でサービスされず、そのサービスのクラスについて指定されたサービスの質に従ってサービスされることを示す。

【0027】ここで図3を参照すると、本発明の論理を見ることができる。図3には、開示を簡単にするために流れのシーケンスで論理が示されているが、下で開示する1つまたは複数のステップは、事象駆動であり、非同期に行われることを理解されたい。

【0028】ブロック36から始めると、システム10のさまざまなノード（すなわち、仮想ディスク（がある場合はそれ）を有するクライアント・コンピュータ12と記憶ノード16）は、たとえばタイム・シンクロナイザ34によって、疎に同期化される。「疎」に同期化されるとは、システム・ノードのクロックが、個々のデータ・アクセス要求の粒度（通常は数十ミリ秒）内で同一の時間基準を有するように同期化されることを意味する。本発明によって認識されるように、このような同期化によって、1つまたは複数のクライアント・コンピュータ12を選好する記憶ノード16が、効果的に選好されないクライアント・コンピュータ12からの要求を無視することがなくなる。

【0029】ブロック38に移って、この論理では、次に、クライアント・コンピュータ12からシステム10の共用記憶装置の記憶ノード16への、たとえば読取要求または書込要求であるデータ・アクセス要求の一部またはすべてと、要求優先順位を相関または他の形で関連付ける。これらの優先順位は、時間に基づく必要はないが、1実施形態では、優先順位が、時間に基づく締切期限によって確立される。データ・アクセス要求の締切期限は、その要求に対する応答の最も遅い所望の時刻を表す。

【0030】いくつかの場合には、データ・アクセス要求の優先順位が、要求自体に付加される。それ以外の状況、たとえば、カーネル・データ構造が優先順位による要求ブロックのタグ付けを許容しない時には、優先順位は、関連するデータ・アクセス要求とは別に送られ、後程要求と照合されて、適当な記憶ノード16に要求と共に送られる。

【0031】好ましい実施形態では、ブロック38で、リアルタイム・アプリケーション・プログラミング・インターフェース（API）を呼び出して、たとえばサポートされるリアルタイム・ストリームのカウントと、使用されるファイル速度およびファイル・フォーマットとに基づいて適当な優先順位を生成する。要求が非リアルタイム要求の場合、すなわち、テキスト・ファイルへのアクセスの要求など、特定の時間期間内の満足を必要としない要求である場合には、NULL締切期限が仮定され、これは、下で完全に説明する形で処理される。

【0032】本発明の意図によれば、APIは、要求元のクライアント・コンピュータ12のエンド・ユーザまたはアプリケーションによって呼び出すことができるが、好ましくは、図1に示されたファイル・システム26など、要求元のエンド・ユーザまたはアプリケーションに関連する中間のサブシステムによって呼び出される。どの場合でも、本発明のAPIは、入出力制御（IOCTL）呼出しとして実装される。好ましいVSD実施態様のための優先順位データ構造体（コメント付き）の例を次に示す。

```
typedef struct timestruct_t DEADLINE;
struct vsd_deadline
{
        char *buf;
                /* データ用のユーザ・バッファ */
        DEADLINE deadline;
                /* 入出力に関連する締切期限 */
        int sector_offset;
                /* 記憶装置のセクタ・オフセット */
        int sector_count;
                /* 読み取るセクタ数 */
};
```

【0033】読取データ・アクセス要求の場合、好ましいAPIは
int ioctl(fd, GIODREAD, &;dead);である。ここで、「fd」は、要求元のアプリケーションまたはファイル・システムから渡されるファイル記述子であり、「GIODREAD」は、大域入出力要求を識別する定数であり、「&;dead」は、関連する優先順位データ構造体を識別する。上に開示した読取IOCTLでは、上に記載したVSD_DEADLINEデータ構造体のフィールドによって指されるバッファにデータが読み取られる。前に述べたように、このデータ構造体には、下で説明する形で使用されるリアルタイム締切期限などの優先順位も含まれる。

【0034】これに対して、書込データ・アクセス要求の場合、好ましいAPIは、
int ioctl(fd, GIODWRITE, &;dead);である。

【0035】上で開示した書込IOCTLでは、上に示したVSD_DEADLINEデータ構造体のフィールドによって指されるバッファからデータが書き込まれる。読取要求の場合と同様に、このデータ構造体には、書込要求に関する優先順位も含まれる。

【００３６】さらに、優先順位は、個々のデータ・アクセス要求が分割されてこれになる可能性があるすべてのサブ要求（パケット）に関連する。具体的に言うと、単一のデータ・アクセス要求が、ネットワーク１４のプロトコルによって決定されるメッセージ量子サイズに基づいて、すべてのパケットが同一の優先順位に関連する複数のパケットに配置される可能性がある。したがって、ブロック４０で、パケット化されたデータ・アクセス要求とそれに関連する優先順位が、ネットワーク１４を介してシステム１０の共用記憶装置の適当な記憶ノード１６に送られる。

【００３７】望むならば、特定のデータ記憶装置２０に対するサブ要求を、関連する記憶コンピュータ１８で再組立することができ、その後、完全な要求としてデータ記憶装置２０に送って、データ記憶装置２０が要求に応答する効率を高めることができる。または、サブ要求を正しいシーケンスで順序付け、適当なデータ記憶装置２０に直接送ることができる。

【００３８】ブロック４０から、この論理はブロック４２に進み、そこで、優先順位を伴うデータ・アクセス要求が、共用記憶装置によって受け取られ、データ・アクセス要求を受け取った記憶ノード１６のそれぞれが、関連する要求待ち行列３２（図１）内で優先順位によって要求を順序付ける。非リアルタイムであり、その結果、この実施形態ではＮＵＬＬ優先順位を割り当てられたデータ・アクセス要求の場合、その要求には、要求待ち行列３２にその非リアルタイム要求が到着した時に、要求待ち行列３２内の最後のリアルタイム・アクセス要求の優先順位（＋１）が割り当てられる。これによって、非リアルタイム要求が、非ＮＵＬＬ締切期限を有する新しい要求に継続的に追い越されることがなくなる。

【００３９】しかし、枯渇しないこと（non-starvation）を保証しながら非リアルタイム要求に対処するための他の方式を使用することができる。たとえば、非リアルタイム要求は、待ち行列化されたリアルタイム要求のすべてをそれぞれの優先順位の範囲内で満足することができる場合、または、非リアルタイム要求に保証される最大待ち時間が満足される場合に限って、満足することができる。

【００４０】図１および図３の簡単な相互参照では、システム・ディスク・コントローラが使用されない場合に、各記憶コンピュータ１８は、その要求待ち行列３２から１時に１つの要求だけを関連するデータ記憶装置２０に渡すことができる。言い換えると、ブロック４２で、要求の満足を制御するのにディスク・コントローラが使用されない時に、同時のデータ・アクセス要求は、データ記憶装置２０に対する１時に１つの入出力動作だけを使用して、順次満足される。これによって、データ記憶装置２０のコントローラが、特定のデータ記憶装置２０の内部の判断基準に基づいて要求の間の順序を変更することがなくなる。したがって、データ・アクセス要求は、データ記憶装置２０の外部の１つまたは複数の検討事項によって定義される順序で応答される。１実施形態では、要求待ち行列３２で、データ記憶装置２０のアクセスのための単一のバッファだけが使用される。

【００４１】単一の記憶コンピュータ１８が複数のデータ記憶装置２０を制御する時には、あるデータ記憶装置２０に対する要求を、その記憶コンピュータ１８によって制御される別のデータ記憶装置２０への要求と同時に行うことができることを理解されたい。このような状況では、記憶コンピュータ１８に、それに関連するデータ記憶装置２０ごとにそれぞれの要求待ち行列が含まれる。

【００４２】判断ブロック４４に移ると、この論理では、ある要求の優先順位を更新するか、要求を打ち切る（「kill」する）かが非同期に判定される。本発明によって認識されるように、要求の優先順位の更新は、要求元のアプリケーションまたはファイル・システムが、オン・デマンドでデータ・ストリームの優先順位を動的にシフトしなければならない時、または、ある時間閾値に達した時、または、追加のリアルタイム・ストリームにオフセットを与えるか、ストリーム呼出しの再スケジューリングを行うため（「停止」、「再開」などのビデオ・カセット・レコーダの機能がサポートされる時）に有用である。さらに、たとえばデータ記憶装置２０が要求に対する応答を停止した時などに未処理の要求を打ち切ると、要求元アプリケーションが、システム１０の共用記憶装置の他のデータ記憶装置２０上の複製データ・ブロックを要求することができるようになる。

【００４３】要求の更新も打切りも行われない場合、この処理は状態４６で終了する。そうでない場合には、この論理はブロック４８に移って、例のＶＳＤ実施形態の場合、以下のように要求の優先順位を更新するか、要求を打ち切る。ＶＳＤのクライアント・コンピュータ１２が、システム１０の共用記憶装置にデータ・アクセス要求を送る時に、そのクライアント・コンピュータ１２は、局所保留中待ち行列に未処理の要求のコピーを維持し、そのコピーにタイマを関連付ける。要求の優先順位を更新するか、要求を打ち切るためには、ブロック４８で、要求がまだ満たされていない場合にはクライアント・コンピュータ１２でその要求に関連付けられたタイマを、強制的に０にする。その後、ブロック５０で、新しい（更新された）優先順位と共にシステム１０の共用記憶装置に要求を再送信するか、打切りの場合には終了要求を送信する。

【００４４】好ましい実施形態では、要求優先順位を更新するＩＯＣＴＬと要求を打ち切るＩＯＣＴＬは次のとおりである。

```
int ioctl(fd,GIODUPDT,&;dead);
int ioctl(fd,GIODKILL,&;dead);ここで、「GIODUPDT」
```

と「GIOCKILL」は、大域入出力要求を識別する定数であり、「&;dead」は、関連する優先順位データ構造体を識別する。

【0045】打切り(「kill」)動作の場合、IOCTLによって、上で開示したVSD_DEADLINEデータ構造体のフィールドによって指されるものと同一のバッファを、保留中の要求が使用するかどうかが判定される。そうである場合には、その要求を行ったプロセスが解放され、その要求は、オペレーティング・システムの共用メモリ・セマンティックスを使用して、少なくとも要求元クライアント・コンピュータの保留中待ち行列から除去される。

【0046】同様に、更新動作の場合、IOCTLによって、上で開示したVSD_DEADLINEデータ構造体のフィールドによって指されるものと同一のバッファを、保留中の要求が使用するかどうかが判定される。そうである場合には、上で述べた、新しい優先順位を有する要求の再送信が、上で述べたように強制される。

【0047】適当なデータ記憶装置20が、更新された要求(または打切り要求)を受け取った時に、そのデータ記憶装置20は、判断ブロック52で、古い要求がまだ関連する要求待ち行列32にあるかどうかを判定する。すなわち、適当なデータ記憶装置20は、古い要求がすでに満足されているかどうかを判定し、そうである場合には、この論理はブロック54に移って、更新された要求または打切り要求を無視する。そうでない場合には、この論理はブロック56に進んで、要求の更新された優先順位に従って(または、要求の打切りと要求待ち行列32からの除去に従って)適当な要求待ち行列32の要求の再順序付けを行い、次いで処理は状態46で終了する。望むならば、「kill」要求について上で述べたように、要求元のクライアント・コンピュータ12でのみ要求を打ち切って、要求元のアプリケーションまたはファイル・システムを解放する必要がある場合がある。打切り要求をリモートのデータ記憶装置20に実際に送信する必要はないと思われる。

【0048】本明細書で図示し、詳細に説明した特定の「コンピュータ・クラスタ用のリアルタイム共用ディスク・システム」は、本発明の上述の目標を完全に達成できるが、これは、本発明の現在好ましい実施形態であり、したがって、本発明で広義に企図する内容を表すものであり、本発明の範囲には、当業者に明白になる他の実施形態が完全に含まれ、したがって、本発明の範囲は、請求項のみによって制限されることを理解されたい。請求項において、特に記載がない限り単数形による要素の言及は「少なくとも1つの」を意味する。

【0049】まとめとして、本発明の構成に関して以下の事項を開示する。

【0050】(1)1つまたは複数のデータ・アクセス要求にそれぞれの優先順位を関連付けるための論理手段と、上記データ・アクセス要求および優先順位を記憶ノードに送るための論理手段と、上記データ・アクセス要求が、それぞれの優先順位を考慮して満足されるように、上記それぞれの優先順位に基づいて上記記憶ノードで上記データ・アクセス要求を順序付けるための論理手段とを含む、1つまたは複数の記憶ノードにデータ・アクセス要求を通信する複数のクライアント・ノードを含むコンピュータ・システム。

(2)更新された優先順位にするために、少なくとも1つのデータ・アクセス要求の優先順位を、記憶ノードによって上記データ・アクセス要求が満たされる前に変更するための論理手段と、上記更新された優先順位に基づいて、上記記憶ノードでデータ・アクセス要求を再順序付けするための論理手段とをさらに含む、上記(1)に記載のシステム。

(3)さらに、少なくとも1つのデータ・アクセス要求を打ち切るための論理手段を含む、上記(1)に記載のシステム。

(4)さらに、計算ノードおよび上記記憶ノードを互いに疎に同期化するための手段を含む、上記(1)に記載のシステム。

(5)各記憶ノードが、少なくとも1つの記憶コンピュータおよび少なくとも1つのデータ記憶装置を含み、上記記憶コンピュータが、上記データ記憶装置が上記データ記憶装置の内部の検討事項に基づいてデータ・アクセス要求に応答するシーケンスを再順序付けすることができなくなるように、上記データ記憶装置に対して1時に1つを超えないデータ・アクセス要求を送るための論理手段を含む、上記(1)に記載のシステム。

(6)上記システムが、仮想共用ディスク・システムである、上記(1)に記載のシステム。

(7)上記優先順位が、時間に基づく締切期限を含む、上記(1)に記載のシステム。

(8)共用記憶システムにデータ・アクセス要求を通信する複数のプロセッサを有するコンピュータ・システムにおいて、データ記憶装置の外部の1つまたは複数の検討事項によって少なくとも部分的に定義される順序で上記データ・アクセス要求に応答するステップを含む、上記共用記憶システムの単一のデータ記憶装置に対する少なくとも2つの同時データ・アクセス要求を満足するための、コンピュータで実施される方法。

(9)上記データ記憶装置の外部の上記1つまたは複数の検討事項が、データ要求優先順位を含む、上記(8)に記載の方法。

(10)上記優先順位が、時間に基づく締切期限を含む、上記(9)に記載の方法。

(11)1つまたは複数の上記データ・アクセス要求にそれぞれの優先順位を関連付けるステップと、各記憶ノードが少なくとも1つのデータ記憶装置を含む、上記共用記憶システム内の記憶ノードに上記データ・アクセス

要求および優先順位を送るステップと、上記データ・アクセス要求が、それぞれの優先順位に従って満足されるように、上記それぞれの優先順位に基づいて上記記憶ノードで上記データ・アクセス要求を順序付けるステップアとをさらに含む、上記（１０）に記載の方法。

（１２）更新された優先順位にするために、少なくとも１つのデータ・アクセス要求の優先順位を、上記データ・アクセス要求が記憶ノードによって満足される前に変更するステップと、上記更新された優先順位に基づいて、上記記憶ノードでデータ・アクセス要求を再順序付けするステップアとをさらに含む、上記（１１）に記載の方法。

（１３）少なくとも１つのデータ・アクセス要求を、上記要求が記憶ノードによって満足される前に打ち切るステップをさらに含む、上記（１０）に記載の方法。

（１４）計算ノードおよび上記記憶ノードを互いに疎に同期化するステップをさらに含む、上記（１０）に記載の方法。

（１５）各記憶ノードが、少なくとも１つの記憶コンピュータおよび少なくとも１つのデータ記憶装置を含み、上記記憶コンピュータが、上記データ記憶装置が上記データ記憶装置の内部の検討事項に基づいてデータ・アクセス要求に対する応答のシーケンスを再順序付けすることができなくなるように、上記データ記憶装置に対して１時に１つを超えないデータ・アクセス要求を送る、上記（１０）に記載の方法。

（１６）上記システムが、仮想共用ディスク・システムである、上記（８）に記載の方法。

（１７）ディジタル処理装置によって読み取ることができるコンピュータ・プログラム記憶装置と、１つまたは複数のデータ・アクセス要求を満足するための方法ステップを実行するために上記ディジタル処理装置によって実行することのできる命令を含む、上記コンピュータ・プログラム記憶装置上のプログラム手段とを含み、上記方法ステップが、上記データ・アクセス要求のうちの少なくとも一部に、それぞれの優先順位を関連付けるステップと、共用記憶システムが上記優先順位を考慮して上記データ・アクセス要求に応答することができるように、上記優先順位および上記データ・アクセス要求を上記共用記憶システムに送るステップとを含む、コンピュータ・プログラム装置。

（１８）上記共用記憶システムが、仮想共用ディスク・システムであり、上記優先順位のうちの少なくとも一部が、時間に基づく締切期限である、上記（１７）に記載のコンピュータ・プログラム装置。

（１９）上記方法ステップが、さらに、少なくとも１つのデータ・アクセス要求の優先順位を、更新された優先順位に基づいて上記共用記憶システム内で上記データ・アクセス要求を再順序付けできるようにするために、上記更新された優先順位にするために上記共用記憶システムによって上記データ・アクセス要求が満足される前に変更するステップを含む、上記（１７）に記載のコンピュータ・プログラム装置。

（２０）上記方法ステップが、さらに、少なくとも１つのデータ・アクセス要求を、上記データ・アクセス要求が上記共用記憶システムによって満足される前に打ち切るステップを含む、上記（１９）に記載のコンピュータ・プログラム装置。

（２１）上記方法ステップが、さらに、上記データ・アクセス要求を互いに疎に同期化するステップを含む、上記（１７）に記載のコンピュータ・プログラム装置。

（２２）ディジタル処理装置によって読み取ることのできるコンピュータ・プログラム記憶装置と、１つまたは複数のデータ・アクセス要求を満足するための方法ステップを実行するために上記ディジタル処理装置によって実行することのできる命令を含む、上記コンピュータ・プログラム記憶装置上のプログラム手段とを含み、上記方法ステップが、それぞれの優先順位に基づいて、上記データ・アクセス要求のうちの少なくとも一部に対して、メモリ・システムを用いて応答するステップを含み、上記優先順位および上記データ・アクセス要求が、上記メモリ・システムに送られる、コンピュータ・プログラム装置。

（２３）上記メモリ・システムが、共用記憶システムであり、上記優先順位が、時間に基づく締切期限であり、上記共用記憶システムが、複数の記憶ノードを含み、各記憶ノードが、少なくとも１つの記憶コンピュータおよび少なくとも１つのデータ記憶装置を含み、上記方法ステップが、上記データ記憶装置の内部の検討事項に基づいてデータ・アクセス要求に対する応答のシーケンスを上記データ記憶装置が再順序付けすることができなくなるように、上記データ記憶装置に対して１時に１つを超えないデータ・アクセス要求を送るステップを含む、上記（２２）に記載のコンピュータ・プログラム装置。

（２４）上記方法ステップが、変更された優先順位メッセージに応答して、上記共用記憶システムによってデータ・アクセス要求が満足される前に、記憶コンピュータ側で上記データ・アクセス要求を再順序付けするステップを含む、上記（２３）に記載のコンピュータ・プログラム装置。

【図面の簡単な説明】

【図１】本発明のシステムを示す概略図である。

【図２】プログラム記憶装置の概略図である。

【図３】本発明の論理を示す流れ図である。

【符号の説明】

１０　システム

１２　クライアント・コンピュータ

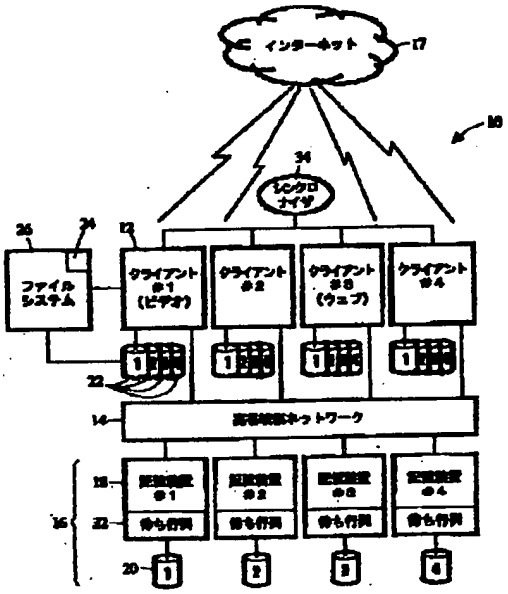
１４　ネットワーク

１６　記憶ノード

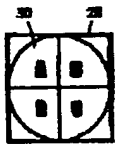
１７　広域ネットワーク（ＷＡＮ）

18 記憶コンピュータ
20 データ記憶装置
22 仮想ディスク
24 締切期限モジュール
26 ファイル・システム
28 コンピュータ・ディスケット
30 コンピュータ使用可能媒体
32 要求待ち行列
34 タイム・シンクロナイザ

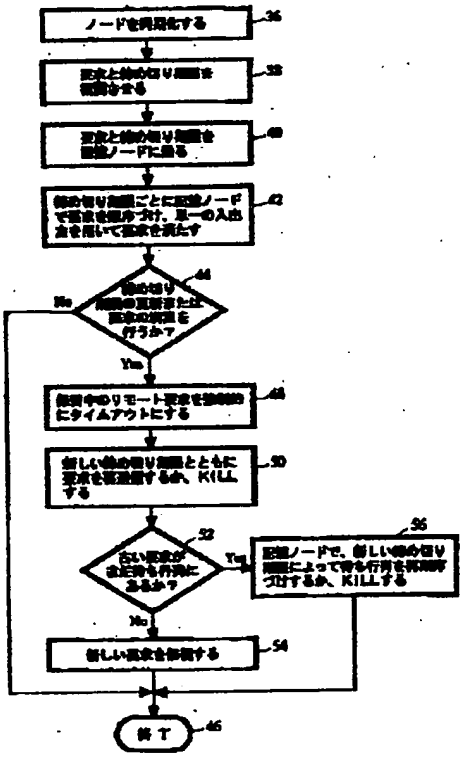
【図1】



【図2】



【図3】



フロントページの続き

(72)発明者 ラージャト・ムケルジー
アメリカ合衆国95136 カリフォルニア州
サンノゼ スネル・アベニュー 4501 アパートメント 3405